AD_____

CONTRACT NO: DAMD17-90-C-0050

TITLE: MOLECULAR STUDIES OF ALPHAVIRUS IMMUNOGENICITY

PRINCIPAL INVESTIGATOR: Dr. James Strauss

CONTRACTING ORGANIZATION: California Institute of Technology
1201 East California Blvd.
Pasadena, California 91125

REPORT DATE: December 3, 1992

TYPE OF REPORT: Annual Report

PREPARED FOR: U.S. Army Medical Research and
Development Command, Fort Detrick
Frederick, Maryland 21702-5012

DISTRIBUTION STATEMENT: Approved for public release;
distribution unlimited

The findings in this report are not to be construed as an official Department of the Army position unless so designated by other authorized documents.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No 0704-0188

1. AGENCY USE ONLY (Leave blank)	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
	3 December 1992	Annual Report (3/30/92 - 9/30/92)

4. TITLE AND SUBTITLE
Molecular Studies of Alphavirus Immunogenicity

5. FUNDING NUMBERS
Contract No.
DAMD17-90-C-0050

61102A
30161102BS12.AB.117
WUDA346101

6. AUTHOR(S)
Dr. James Strauss

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
California Institute of Technology
1201 East California Blvd.
Pasadena, California 91125

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
U.S. Army Medical Research & Development Command
Fort Detrick
Frederick, Maryland 21702-5012

10. SPONSORING/MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES

12a. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

12b. DISTRIBUTION CODE

13. ABSTRACT *(Maximum 200 words)*

In the past 6 months we have completed the sequences of Whataroa virus and Aura virus, and these sequences are being assembled into the complete sequences of the genomic RNAs. These two viruses are both closely related to Sindbis virus, and Sindbis-like viruses have now been shown to occur throughout the temperate and tropical world except for North America. Aura virus may have served as the ancestral virus that was transmitted to the Old World to found the Sindbis-like viruses present throughout the Old World. It may also have served as one of the parents of Western equine encephalitis virus, which arose by recombination between Eastern equine encephalitis virus and a New World virus related to Sindbis virus. In this project we have also developed methods to use high throughout automated sequencing in order to rapidly obtain sequence data for entire viral RNA genomes.

14. SUBJECT TERMS
Alphavirus/Sindbis virus/Whataroa virus/Aura virus
BD/Epitopes/Immunogenetics/DNA/BL2/Biotechnology/
RA 4

15. NUMBER OF PAGES

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
Unclassified	Unclassified	Unclassified	Unlimited

FOREWORD

Opinions, interpretations, conclusions and recommendations are those of the author and are not necessarily endorsed by the US Army.

_____ Where copyrighted material is quoted, permission has been obtained to use such material.

_____ Where material from documents designated for limited distribution is quoted, permission has been obtained to use the material.

_____ Citations of commercial organizations and trade names in this report do not constitute an official Department of Army endorsement or approval of the products or services of these organizations.

_____ In conducting research using animals, the investigator(s) adhered to the "Guide for the Care and Use of Laboratory Animals," prepared by the Committee on Care and Use of Laboratory Animals of the Institute of Laboratory Resources, National Research Council (NIH Publication No. 86-23, Revised 1985).

____ For the protection of human subjects, the investigator(s) adhered to policies of applicable Federal Law 45 CFR 46.

JAS In conducting research utilizing recombinant DNA technology, the investigator(s) adhered to current guidelines promulgated by the National Institutes of Health.

12-3-90 JH Stevens

PI - Signature DATE

Table of Contents

	Page
Front Cover	1
Foreword	2
Report Documentation Page	3
Table of Contents	4
Report	5
Introduction	5
Methods Used	6
Sequence Analysis of Whataroa Virus	7
Sequence of Aura Virus	11
Conclusions	21
References	22

Introduction

The alphaviruses are a widespread group of human pathogens that are present virtually everywhere in the world (Griffin, 1986; Monath, 1988 #1774; Peters, 1990 #1551). They are mosquito-borne viruses and have the capacity to replicate in the mosquito vector as well as in human host or in various species of birds and mammals. Old World alphaviruses are, in general, capable of causing fever, rash and arthralgia in man that may be very painful and disabling for extended periods of time. In the case of the Ockelbo strain of Sindbis virus and of Ross River virus, this arthralgia manifests as a polyarthritis that may in some cases last for months or years. Many of the New World alphaviruses can cause fatal encephalitis in man. Our program attempts to understand the molecular basis of alphaviruses immunogenicity and determine the relationships of alphaviruses and strains of alphaviruses to one another.

In our last report we reported the localization of a site in alphavirus glycoprotein E2 that binds neutralizing antibodies. The knowledge of immunogenic domains is important in developing vaccines. Neutralizing antibodies are thought to be particularly important in protecting a vaccinee from viral infection. We developed a novel approach in which λgt11 expression libraries were constructed that expressed parts of the Sindbis genome, and these were screened with neutralizing monoclonal antibodies. Many neutralizing antibodies react with discontinuous epitopes and thus will not react with a chimeric protein expressed in a λgt11 library. However, we did succeed in identifying one antibody which bound to specific clones within the λgt11 library (Wang and Strauss, 1991). Thus we were able to demonstrate directly that this neutralizing monoclonal antibody bound to glycoprotein E2 of Sindbis virus between residues 173 and 220. This approach confirmed and extended results in which variants of the virus selected to be resistant to neutralizing monoclonal

antibodies were sequenced in order to identify the regions within the glycoproteins of the virus with which the antibodies react (Strauss et al., 1991). We thus identified the domain between residues 170 and 220 of glycoprotein E2 of alphaviruses as being particularly important for the antibody response of a host.

We have also reported on the sequence analysis of a number of strains of Sindbis virus or of viruses related to Sindbis virus, in order to understand the relationships of these viruses to one another. We found that a strain of Sindbis virus from Northern Europe that causes Ockelbo disease in Sweden, Pogosta disease in Finland, or Karelian fever in Russia, a disease characterized by a polyarthritis whose symptoms can persist for months or years, are very closely related to pathogenic strains of Sindbis virus isolated from South Africa. We concluded that a South African strain of Sindbis was introduced into Northern Europe, probably in the 1960s, where it continues to cause epidemics of a significant human disease (Shirako et al., 1991). We have also reported on sequences of a number of other Sindbis-like viruses in order to determine the relationships of these viruses to one another. In this report, we present sequence data for a Sindbis-like virus isolated from New Zealand, Whataroa virus, and a virus from South America, Aura virus, which has been isolated from Brazil and from Argentina. We have been particularly interested in Aura virus because it might represent the parent of an emergent virus, Western equine encephalitis virus.

Methods Used

Virus Strains. Whataroa virus and Aura virus were obtained from Dr. J. M. Dalrymple of USAMRIID. Viruses were grown and purified as previously described (Shirako et al., 1991).

cDNA Clones. cDNA clones were made in one of two ways. The first method used standard procedures in which first strand cDNA was made using oligo(dT) as primer and second strand synthesis was by the method of Gubler and Hoffman (Sambrook et al., 1989); Gubler, 1983 #1546. These cloning methods, as well as the methods of DNA sequencing and RNA sequencing, have been described in numerous publications from our laboratory over the years (Hahn et al., 1985; Rice et al., 1985; Rice and Strauss, 1981; Shirako et al., 1991; Strauss et al., 1984).

In a second approach, we developed methods suitable for high throughput automated DNA sequencing, in order to speed up the acquisition of sequence data. Whataroa virus was chosen as a test virus. First strand cDNA synthesis used random priming and second strand cDNA was synthesized by the method of Gubler and Hoffman (Gubler and Hoffman, 1983). After blunt ending the double-stranded cDNA, the internal EcoRI sites were methylated and the DNA was electrophoresed in an agarose gel. EcoRI linkers were attached to the 2-4 kb fraction and the DNA cloned in the EcoRI site of a suitable vector. One hundred clones that resulted from this cloning were characterized by restriction analysis and many of them were sequenced using an Applied Biosystems automated DNA sequencer.

Sequence Analysis of Whataroa Virus.

In our report of April 24th of this year, we reported the sequence of nsP3 and of nsP4 of Whataroa virus. Most of the sequence of Whataroa virus RNA, 11.7 kb, has now been obtained. This sequence is being assembled to give the complete sequence of this virus RNA. The sequences of two stretches of the nonstructural protein coding region of the genome are shown in Figs. 1 and 2 as an example of this assembly process. Fig. 1 shows the sequence of about 1000 nucleotides

```
1          AAA CAG CCG ACC AAT TGC ACT ACC ATC ACT ATG GAG AAG CCC GTT GTC AAC GTA GAC GTA
                                                   met glu lys pro val val asn val asp val

61/11      GAC CCT CAA AGT CCG TTC GTT GCA CAA CTG CAG AAG AGC TTC CCT CAA TTT GAG GTA GTT
           asp pro gln ser pro phe val ala gln leu gln lys ser phe pro gln phe glu val val

121/31     GCC CAG CAG GCC ACG CCA AAT GAC CAT GCT AAT GCC AGA GCC TTT TCG CAT CTG GCT AGT
           ala gln gln ala thr pro asn asp his ala asn ala arg ala phe ser his leu ala ser

181/51     AAA CTG ATC GAG CTG GAG GTG CCT ACC ACA GCG ACG ATC TTG GAC ATC GGC AGC GCA CCT
           lys leu ile glu leu glu val pro thr thr ala thr ile leu asp ile gly ser ala pro

241/71     GCT CGT AGA ATG TTT TCC GAG CAC CAA TAC CAT TGC GTC TGC CCC ATG CGT AGT CCC GAA
           ala arg arg met phe ser glu his gln tyr his cys val cys pro met arg ser pro glu

301/91     GAC CCG GAC CGC ATG ATG AAA TAC GCC GCC AAA CTG GCA GAA AAA GCA GGA TCT TTA ACC
           asp pro asp arg met met lys tyr ala ala lys leu ala glu lys ala gly ser leu thr

361/111    AAC AAA AAG TTG TAC GAA AAG ATC CGC GAC TTA AGA ACC GTT CTG GAC ACT CCA GAC CAA
           asn lys lys leu tyr glu lys ile arg asp leu arg thr val leu asp thr pro asp gln

421/131    GAA ACA CCA TCC ATA TGC TTC CAT AAC GAC GTA ACC TGC GCT ACA CGA GCA GAA GTA TCG
           glu thr pro ser ile cys phe his asn asp val thr cys ala thr arg ala glu val ser

481/151    GTA ATG CAA GAC GTG TAC ATC AAT GCA CCT GCC ACC ATC TAC CAT CAG GCA ATG AAA GGA
           val met gln asp val tyr ile asn ala pro ala thr ile tyr his gln ala met lys gly

541/171    GTT CGC ACG CTC TAT TGG ATT GGG TTC GAC ACC ACT CAA TTC ATG TTC TCG GCC ATG GCA
           val arg thr leu tyr trp ile gly phe asp thr thr gln phe met phe ser ala met ala

601/191    GGG TCC TAC CCC GCT TAC AAC ACC AAT TGG GCA GAC GAG AAA GTA CTC GAA GCC AGA AAC
           gly ser tyr pro ala tyr asn thr asn trp ala asp glu lys val leu glu ala arg asn

661/211    ATT GGA CTA TGC AGC ACA AAG TTA AGC GAG GGG AGG TTG GGG AAA CTT TCG ATC ATG AGG
           ile gly leu cys ser thr lys leu ser glu gly arg leu gly lys leu ser ile met arg

721/231    AAG AAG TCA TTG AAG CCT GGG ACC CAG GTT TAT TTT TCA GTT GGT TCG ACG TTG TAC CCC
           lys lys ser leu lys pro gly thr gln val tyr phe ser val gly ser thr leu tyr pro

781/251    GAA AAC CGC GCC AAC TTG CAA AGT TGG CAT TTG CCA TCT GTT TTT CAT CTG AAA GGC AAG
           glu asn arg ala asn leu gln ser trp his leu pro ser val phe his leu lys gly lys

841/271    CAA CCA TAC ACC TGC CGC TGT GAT ACA GTG GTA AGC TGT GAA GGC TAC GTA GTC AAG AAA
           gln pro tyr thr cys arg cys asp thr val val ser cys glu gly tyr val val lys lys

901/291    GTG ACT ATC AGT CCC GGG ATA ACC GGA GAA ACC GTG GGA TAC GCG GTG ACT AAC AAC AGT
           val thr ile ser pro gly ile thr gly glu thr val gly tyr ala val thr asn asn ser

961/311    GAG GGA TTC TTG CTG TGC AAA GTC ACA GAC ACA GTA AAA GGG GAA CGG GTC TCG TTT CCC
           glu gly phe leu leu cys lys val thr asp thr val lys gly glu arg val ser phe pro

1021/331   GTA TGT ACT TAC ATA CCA GCT ACT ATC TGT GAC CAA ATG ACT GGG ATC ATG
           val cys thr tyr ile pro ala thr ile cys asp gln met thr gly ile met
```

Figure 1.Translated nucleotide sequence from the 5'terminal region of the genomic RNA of Whataroa virus, using the single letter amino acid code. The open reading frame begins with the ATG codon (nt 31-33). The exact 5' terminus of the RNA has not been determined.

```
   1   F  I  N  R  K  L  Y  H  I  A  V  H  G  P  A  K  N  T  E  E      20
   1   TTCATTAACAGGAAATTGTACCACATTGCAGTTCATGGTCCCGCGAAGAATACTGAGGAA    60
                |        |        |        |        |
  21   E  Q  Y  K  A  M  R  A  E  A  A  D  T  E  Y  V  F  D  V  D      40
  61   GAGCAGTATAAAGCTATGAGAGCAGAAGCGGCGGACACCGAATATGTCTTCGATGTCGAC    120
                |        |        |        |        |
  41   K  K  K  C  V  K  R  E  E  A  S  G  L  V  L  V  G  E  L  T      60
 121   AAGAAGAAGTGCGTTAAGAGAGAAGAAGCATCGGGTCTTGTGTTAGTAGGCGAACTTACC    180
                |        |        |        |        |
  61   N  P  P  Y  H  E  M  A  L  E  G  L  K  T  R  P  A  V  P  Y      80
 181   AACCCGCCATACCATGAAATGGCGCTGGAAGGGCTGAAGACCCGTCCTGCAGTACCTTAT    240
                |        |        |        |        |
  81   K  V  E  T  I  G  V  I  G  T  P  G  S  G  K  S  A  I  I  K     100
 241   AAAGTTGAAACAATCGGAGTCATCGGCACACCGGGATCCGGAAAATCCGCAATCATTAAA    300
                |        |        |        |        |
 101   N  I  V  T  T  R  D  L  V  T  S  G  K  K  E  N  C  R  E  I     120
 301   AACATCGTCACTACCAGGGATCTTGTGACCAGCGGAAAGAAAGAAAACTGCCGGGAAATA    360
                |        |        |        |        |
 121   E  A  D  V  L  K  H  R  K  M  Q  I  V  S  K  T  V  D  S  V     140
 361   GAAGCTGACGTCCTCAAACACCGAAAAATGCAAATCGTTTCAAAGACGGTCGACTCCGTT    420
                |        |        |        |        |
 141   L  L  N  G  C  H  K  S  V  D  I  L  Y  V  D  E  A  Y  A  C     160
 421   TTGCTTAATGGTTGCCACAAGTCAGTCGACATCCTGTATGTCGACGAAGCTTACGCGTGC    480
                |        |        |        |        |
 161   H  A  G  T  L  L  A  L  I  A  I  V  R  P  R  N  K  V  V  L     180
 481   CACGCTGGCACCCTATTGGCCTTAATCGCCATAGTCCGACCTAGAAATAAAGTGGTCCTA    540
                |        |        |        |        |
 181   C  G  D  P  K  Q  C  G  F  F  N  M  M  Q  L  K  V  H  F  N     200
 541   TGTGGCGACCCAAAACAGTGTGGTTTCTTCAACATGATGCAGCTGAAGGTCCACTTTAAC    600
                |        |        |        |        |
 201   D  P  E  R  D  I  C  T  K  T  F  Y  K  Y  I  S  R  R  C  T     220
 601   GACCCTGAACGCGACATTTGCACGAAGACGTTCTACAAATACATTTCTCGTCGGTGCACG    660
                |        |        |        |        |
 221   Q  P  V  T  A  I  V  S  T  L  H  Y  N  G  K  M  R  T  T  N     240
 661   CAACCGGTGACAGCAATTGTGTCTACACTGCACTATAACGGAAAAATGCGCACCACCAAC    720
                |        |        |        |        |
 241   P  C  N  K  N  I  V  I  D  I  T  G  Q  T  K  P  K  P  G  D     260
 721   CCATGTAACAAGAACATCGTAATCGACATTACCGGACAAACCAAACCAAAACCAGGAGAT    780
                |        |        |        |        |
 261   I  I  L  T  C  F  R  G  W  V  K  Q  L  Q  I  E  Y  P  G  H     280
 781   ATTATCCTGACGTGTTTCAGGGGGTGGGTCAAGCAGCTGCAGATTGAATACCCAGGACAC    840
                |        |        |        |        |
 281   E  V  M  T  A  A  V  S  Q  G  L  T  R  K  G  V  F  P  V  R     300
 841   GAAGTTATGACTGCGGCAGTTTCACAAGGATTGACGCGAAAAGGGGTCTTTCCCGTAAGA    900
                |        |        |        |        |
 301   G  K  V  N  E  N  P  L  Y  A  I  T  S  E  H  V  N  V  L  L     320
 901   GGAAAAGTCAACGAGAACCCGTTATATGCCATCACTTCTGAGCACGTCAACGTACTGTTG    960
                |        |        |        |        |
 321   T  R  T  E  D  R  I  V  W  K  T  L  Q  G  D  P  W  I  K  Q     340
 961   ACACGAACCGAAGATCGTATCGTGTGGAAAACGCTACAAGGAGACCCTTGGATAAAGCAG   1020
                |        |        |        |        |
 341   L  T  N  I  P  K  G  N  F  H  A  T  V  E  E  W  E  A  E  H     360
1021   CTCACAAACATTCCAAAAGGCAACTTTCACGCCACCGTCGAAGAATGGGAGGCTGAACAC   1080
                |        |        |        |        |
```

Figure 2 See legend on next page.

```
 361  K  G  I  M  E  A  I  T  S  P  A  P  R  S  N  P  F  S  C  K      380
1081  AAGGGAATAATGGAGGCTATCACTAGCCCGGCCCCCCGCAGCAACCCTTTCAGCTGTAAG  1140

 381  T  N  V  C  W  A  K  A  L  E  P  I  L  S  T  A  G  I  S  L      400
1141  ACAAACGTGTGCTGGGCGAAGGCACTAGAACCTATACTATCGACCGCTGGCATATCACTA  1200

 401  T  G  C  Q  W  A  D  L  F  P  Q  F  E  D  D  K  P  H  S  A      420
1201  ACTGGATGTCAGTGGGCAGATTTGTTTCCGCAATTTGAAGATGACAAACCACATTCGGCC  1260

 421  I  Y  A  L  D  V  I  C  V  K  F  F  G  M  D  L  T  S  G  I      440
1261  ATATACGCTCTAGACGTCATTTGCGTAAAGTTCTTTGGCATGGATTTAACTAGCGGCATA  1320

 441  F  S  K  P  L  I  P  L  T  Y  H  P  A  E  G  D  R  K  T  A      460
1321  TTTTCAAAACCGTTGATCCCATTGACTTATCACCCCGCCGAAGGGGACCGGAAGACAGCG  1380

 461  H  W  D  N  S  P  G  Q  R  K  Y  G  F  D  K  A  V  V  A  E      480
1381  CACTGGGACAACAGTCCAGGCCAACGAAAGTACGGGTTTGACAAAGCCGTTGTAGCTGAA  1440

 481  L  S  R  R  F  P  V  F  C  M  A  D  K  G  V  Q  L  D  L  Q      500
1441  TTGTCCCGCAGATTCCCAGTATTCTGCATGGCAGACAAAGGAGTGCAACTGGACCTACAG  1500

 501  T  G  R  T  R  V  V  ?  S  R  F  N  L  V  P  F  N  R  N  L      520
1501  ACGGGCCGNACGCGCGTAGTCNCGTCACGCTTCAACCTTGTGCCATTTAACAGAAATCTG  1560

 521  P  H  S  L  V  P  E  Y  K  T  Q  T  P  G  Q  L  S  A  F  I      540
1561  CCCCACTCGCTTGTCCCGGAGTATAAAACACAAACTCCAGGTCAGCTAAGCGCCTTTATC  1620

 541  R  Q  F  K  Q  N  T  I  L  L  V  S  E  T  P  A  E  H  S  T      560
1621  CGCCAGTTTAAACAAAACACCATCCTGCTTGTATCTGAAACACCTGCCGAACATTCCACC  1680

 561  K  S  V  E  W  I  A  P  L  G  T  L  G  A  T  K  C  Y  N  L      580
1681  AAATCTGTGGAATGGATTGCACCGCTGGGTACGCTTGGAGCCACCAAATGCTATAATTTA  1740

 581  A  F  G  F  P  P  Q  S  R  Y  D  L  V  I  I  N  I  G  T  K      600
1741  GCATTCGGCTTTCCGCCTCAGTCGAGGTACGACCTAGTGATCATAAATATCGGTACAAAA  1800

 601  F  R  H  H  H  Y  Q  Q  C  E  D  H  A  A  T  M  K  T  L  S      620
1801  TTCAGACACCACCACTATCAACAGTGCGAAGACCACGCCGCCACCATGAAGACACTGTCA  1860

 621  R  S  A  L  N  C  L  N  P  G  G  T  L  V  V  K  A  Y  G  Y      640
1861  CGTTCCGCCCTTAATTGCCTGAACCCGGGTGGCACATTGGTGGTAAAAGCATATGGCTAC  1920

 641  A  D  R  N  S  E  D  I  I  T  A  L  A  R  K  F  V  R  V  S      660
1921  GCGGACAGAAACAGTGAAGACATCATTACAGCCCTGGCACGAAAGTTCGTCAGGGTGTCC  1980

 661  A  A  R  P  Q  C  V  S  S  N  T  E  M  Y  F  I  F  R  Q  L      680
1981  GCGGCCCGCCCACAGTGCGTCTCAAGCAATACAGAGATGTACTTCATTTTCAGACAACTG  2040

 681  D  N  S  R  T  R  Q  F  T  P  H  H  L  N  C  V  V  S  S  V      700
2041  GACAACAGCAGAACACGTCAATTCACACCTCATCACCTCAACTGCGTCGTTTCGTCAGTG  2100

 701  Y  E  G  T  R  D  G  V  G  A                                    710
2101  TACGAGGGAACAAGAGACGGAGTTGGTGCT                                  2130
```

Figure 2 continued. Translated nucleotide sequence of Whataroa virus in the region encoding nonstructural protein nsP2. By homology with Sindbis virus, the sequence shown begins at amino acid 97 of nsP2 and continues to the nsP2/nsP3 cleavage site.

beginning in the 5' nontranslated region just upstream of the start codon of the long open reading frame translated from the viral genomic RNA. The second sequence of about 2000 nucloeotides begins near the beginning of the nsP2 gene and continues through to the end of the nsP2 region of the virus genome. As stated, the remainder of the sequence has been obtained and is being assembled.

Whataroa virus can clearly be considered to be a strain of Sindbis virus that has spread to New Zealand. The amino acid sequence deduced from the nucleotide sequence in Fig. 2 is compared to that of the AR339 strain of Sindbis virus, isolated from Egypt in 1952, in Fig. 3. These amino acid sequences are 84% identical. Furthermore, we have previously shown that strains of Sindbis virus contain a 3' nontranslated regions that is different from all other alphaviruses. It contains three copies of a sequence that is conserved among Sindbis viruses that are spaced by sequences that are poorly conseved (Shirako et al., 1991). From our sequence data, we found that this characteristic 3' nontranslated region is present in Whataroa virus.

Sequence of Aura Virus.

The sequence of essentially all of the Aura virus genome has also been obtained and is being assembled. As an example of this assembly process, the sequence of about 5000 nucleotides of Aura RNA in the nonstructural protein coding region is shown in Fig. 4. This sequence begins in the 5' nontranslated region and continues through nsP1, nsP2, and part of nsP3. Aura virus is closely related to Sindbis virus. The amino acid sequences of Sindbis virus and of Aura virus are compared in Fig. 5 for the region represented by the Aura sequence in Fig. 4. The two sequences are 80% identical, illustrating that Aura is in fact a Sindbis-like virus. We also found that the 3' nontranslated region of Aura RNA is Sindbis-like. Thus Aura virus represents the first known example of a true

```
FINRKLYHIAVHGPAKNTEEEQYKAMRAEAADTEYVFDVDKKKCVKREEA
.V........M.............VTK..L.E..........R...K...
         *          *          *          *          *
SGLVLVGELTNPPYHEMALEGLKTRPAVPYKVETIGVIGTPGSGKSAIIK
.....S..........L..................................
         *          *          *          *          *
NIVTTRDLVTSGKKENCREIEADVLKHRKMQIVSKTVDSVLLNGCHKSVD
ST..A.....................RL.G...T.......M .....A.E
         *          *          *          *          *
ILYVDEAYACHAGTLLALIAIVRPRNKVVLCGDPKQCGFFNMMQLKVHFN
V......F.....A...........K........M................
         *          *          *          *          *
DPERDICTKTFYKYISRRCTQPVTAIVSTLHYNGKMRTTNPCNKNIVIDI
H..K.............................D...K.....K...E...
         *          *          *          *          *
TGQTKPKPGDIILTCFRGWVKQLQIEYPGHEVMTAAVSQGLTRKGVFPVR
..A........................D..........A..........YA..
         *          *          *          *          *
GKVNENPLYAITSEHVNVLLTRTEDRIVWKTLQGDPWIKQLTNIPKGNFH
Q............................L.............P........Q
         *          *          *          *          *
ATVEEWEAEHKGIMEAITSPAPRSNPFSCKTNVCWAKALEPILSTAGISL
..I.D........IA..N..T..A.....................A....V.
         *          *          *          *          *
TGCQWADLFPQFEDDKPHSAIYALDVICVKFFGMDLTSGIFSKPLIPLTY
.....SE.....A................I..........L...QS.....
         *          *          *          *          *
HPAEGDRKTAHWDNSPGQRKYGFDKAVVAELSRRFPVFCMADKGVQLDLQ
...DSA.PV........T....Y.H.IA..........QL.G..T......
         *          *          *          *          *
TGRTRVV?SRFNLVPFNRNLPHSLVPEYKTQTPGQLSAFIRQFKQNTILL
......ISAQH....V......A......EKQ..PVKK.LN...HHSV.V
         *          *          *          *          *
VSETPAEHSTKSVEWIAPLGTLGATKCYNLAFGFPPQSRYDLVIINIGTK
...EKI.APR.RI.....I.IA..D.N..........A.....F......
         *          *          *          *          *
FRHHHYQQCEDHAATMKTLSRSALNCLNPGGTLVVKAYGYADRNSEDIIT
Y.N..F.........L.....................S..........VV.
         *          *          *          *          *
ALARKFVRVSAARPQCVSSNTEMYFIFRQLDNSRTRQFTPHHLNCVVSSV
..............D.........L.....................I...
         *          *          *          *          *
YEGTRDGVGA
..........
```

Figure 3. Aligned deduced amino acid sequences of the nonstructural protein regions of Whataroa virus and Sindbis virus, beginning with amino acid 97 of Sindbis virus nsP2. The upper sequence in each case is Whataroa virus, and amino acid identity in the Sindbis sequence is indicated with a dot.

```
1          ACT AGT ACT TGT ACT ACA GAA TTA ACT GCC GTG TGC CGC CCG CTA AAC TAG CCC CAA TCA

61         TCG AAA ATG GAG AAA CCG ACA GTG CAC GTT GAC GTA GAC CCC CAA AGT CCG TTT GTG CTA
                  met glu lys pro thr val his val asp val asp pro gln ser pro phe val leu

121/19     CAA CTG CAG AAG AGT TTC CCA CAA TTC GAG ATT GTG GCT CAG CAG GTC ACT CCG AAT GAC
           gln leu gln lys ser phe pro gln phe glu ile val ala gln gln val thr pro asn asp

181/39     CAT GCT AAT GCC AGA GCT TTT TCG CAT CTG GCT AGT AAA CTG ATC GAA CAT GAG ATC CCC
           his ala asn ala arg ala phe ser his leu ala ser lys leu ile glu his glu ile pro

241/59     ACC TCA GTT ACG ATC TTG GAC ATA GGA AGC GCA CCA GCT CGT AGA ATG TAT TCC GAG CAT
           thr ser val thr ile leu asp ile gly ser ala pro ala arg arg met tyr ser glu his

301/79     AAG TAT CAC TGT GTG TGC CCC ATG CGT AGT CCT GAA GAC CCG GAC CGT CTT ATG AAT TAC
           lys tyr his cys val cys pro met arg ser pro glu asp pro asp arg leu met asn tyr

361/99     GCA TCC CGA CTC GCA GAC AAA GCA GGG GAA ATT ACC AAC AAG AGG CTG CAT GAT AAA CTT
           ala ser arg leu ala asp lys ala gly glu ile thr asn lys arg leu his asp lys leu

421/119    GCA GAC CTC AAG TCG GTC CTC GAG TCG CCG GAT GCT GAA ACT GGT ACC ATT TGT TTC CAC
           ala asp leu lys ser val leu glu ser pro asp ala glu thr gly thr ile cys phe his

481/139    AAT GAC GTA ATA TGC CGT ACG ACA GCG GAG GTA TCA GTT ATG CAA AAT GTG TAT ATC AAT
           asn asp val ile cys arg thr thr ala glu val ser val met gln asn val tyr ile asn

541/159    GCA CCT TCG ACC ATT TAC CAT CAG GCC CTA AAG GGA GTC AGA AAA CTG TAT TGG ATC GGG
           ala pro ser thr ile tyr his gln ala leu lys gly val arg lys leu tyr trp ile gly

601/179    TTC GAT ACA ACG CAG TTT ATG TTC TCC TCG ATG GCA GGG TCG TAT CCG TCC TAC AAT ACT
           phe asp thr thr gln phe met phe ser ser met ala gly ser tyr pro ser tyr asn thr

661/199    AAT TGG GCC GAT GAA AGG GTG CTG GAA GCG CGT AAT ATA GGC CTA TGT AGC ACG AAG CTG
           asn trp ala asp glu arg val leu glu ala arg asn ile gly leu cys ser thr lys leu

721/219    AGA GAG GGT ACG ATG GGC AAA CTG TCT ACC TTC CGG AAA AAG GCC TTG AAA CCT GGA ACT
           arg glu gly thr met gly lys leu ser thr phe arg lys lys ala leu lys pro gly thr

781/239    AAC GTG TAC TTC TCT GTC GGT TCG ACA CTC TAC CCT GAG AAT AGA GCG GAC CTG CAG AGT
           asn val tyr phe ser val gly ser thr leu tyr pro glu asn arg ala asp leu gln ser

841/259    TGG CAC CTA CCA TCT GTG TTC CAC TTG AAA GGT AAA CAA TCC TTT ACG TGC CGC TGT GAT
           trp his leu pro ser val phe his leu lys gly lys gln ser phe thr cys arg cys asp

901/279    ACG GCG GTT AAC TGC GAA GGA TAC GTA GTC AAG AAG ATC ACC ATC AGC CCC GGG ATC ACG
           thr ala val asn cys glu gly tyr val val lys lys ile thr ile ser pro gly ile thr

961/299    GGG CGT GTC AAT CGG TAC ACT GTG ACT AAC AAC AGC GAG GGA TTC TTG CTG TGT AAG ATC
           gly arg val asn arg tyr thr val thr asn asn ser glu gly phe leu leu cys lys ile

1021/319   ACA GAT ACG GTC AAA GGG GAG CGT GTA TCG TTC CCT GTC TGT ACG TAT ATT CCA CCT TCA
           thr asp thr val lys gly glu arg val ser phe pro val cys thr tyr ile pro pro ser

1081/339   ATC TGT GAC CAA ATG ACA GGT ATA TTG GCC ACT GAT ATC CAA CCC GAA GAC GCG CAA AAG
           ile cys asp gln met thr gly ile leu ala thr asp ile gln pro glu asp ala gln lys
```

Figure 4a. See legend on last page of this sequence

```
1141/359  TTG CTG GTA GGA CTG AAC CAA CGC ATA GTC GTG AAC GGA AAA ACT AAT AGA AAC ACC AAC
          leu leu val gly leu asn gln arg ile val val asn gly lys thr asn arg asn thr asn

1201/379  ACG ATG CAG AAC TAT CTC CTG CCC GCG GTG GCT ACA GGT CTG AGT AAA TGG GCC AAA GAA
          thr met gln asn tyr leu leu pro ala val ala thr gly leu ser lys trp ala lys glu

1261/399  AGA AAG GCA GAC TGC AGT GAC GAG AAA CCA TTG AAT GTG AGA GAA CGC AAA CTA GCT TTC
          arg lys ala asp cys ser asp glu lys pro leu asn val arg glu arg lys leu ala phe

1321/419  GGT TGC CTA TGG GCT TTC AAG ACC AAG AAG ATC CAT TCT TTT TAC CGC CCG CCA GGC ACG
          gly cys leu trp ala phe lys thr lys lys ile his ser phe tyr arg pro pro gly thr

1381/439  CAG ACT ATA GTA AAA GTC GCA GCG GAA TTC AGT GCG TTC CCT ATG TCC TCG GTG TGG ACT
          gln thr ile val lys val ala ala glu phe ser ala phe pro met ser ser val trp thr

1441/459  ACG TCA CTG CCA ATG TCA CTG AGA CAG AAA GTT AAA CTG CTT CTT GTA AAG AAA ACC AAT
          thr ser leu pro met ser leu arg gln lys val lys leu leu leu val lys lys thr asn

1501/479  AAA CCG GTA GTC ACT ATT ACT GAC ACT GCG GTA AAA AAC GCA CAA GAG GCA TAT AAC GAA
          lys pro val val thr ile thr asp thr ala val lys asn ala gln glu ala tyr asn glu

1561/499  GCC GTC GAG ACA GCA GAA GCG GAG GAG AAA GCG AAG GCC TTA CCT CCG CTG AAG CCG ACG
          ala val glu thr ala glu ala glu glu lys ala lys ala leu pro pro leu lys pro thr

1621/519  GCA CCC CCT GTA GCG GAG GAC GTC AAA TGC GAG GTC ACC GAC CTG GTA GAC GAT GCG GGA
          ala pro pro val ala glu asp val lys cys glu val thr asp leu val asp asp ala gly

1681/539  GCG GCC CTG GTC GAG ACG CCC CGG GGA AAG ATA AAA ATT ATC CCA CAG GAA GGG GAC GTG
          ala ala leu val glu thr pro arg gly lys ile lys ile ile pro gln glu gly asp val

1741/559  CGT ATT GGT TCC TAC ACA GTC ATT TCT CCA GCG GCA GTC CTT AGA AAT CAA CAA CTG GAG
          arg ile gly ser tyr thr val ile ser pro ala ala val leu arg asn gln gln leu glu

1801/579  CCA ATC CAC GAG TTA GCA GAG CAG GTG AAA ATT ATC ACG CAC GGT GGC CGA ACA GGC AGG
          pro ile his glu leu ala glu gln val lys ile ile thr his gly gly arg thr gly arg

1861/599  TAT TCC GTC GAA CCT TAC GAT GCT AAG GTT CTC CTG CCA ACA GGA TGC CCC ATG TCC TGG
          tyr ser val glu pro tyr asp ala lys val leu leu pro thr gly cys pro met ser trp

1921/619  CAA CAT TTC GCG GCC TTG AGC GAA AGC GCT ACG TTA GTC TAC AAT GAG AGA GAG TTC CTG
          gln his phe ala ala leu ser glu ser ala thr leu val tyr asn glu arg glu phe leu

1981/639  AAC CGG AAA CTC CAT CAC ATC GCT ACG AAG GGT GCG GCA AAA AAC ACT GAG GAA GAA CAA
          asn arg lys leu his his ile ala thr lys gly ala ala lys asn thr glu glu glu gln

2041/659  TAC AAA GTA TGC AAA GCT AAA GAC ACG GAT CAT GAG TAC GTA TAC GAC GTA GAT GCC AGA
          tyr lys val cys lys ala lys asp thr asp his glu tyr val tyr asp val asp ala arg

2101/679  AAA TGC GTA AAA AGA GAG CAT GCA CAA GGG CTA GTA CTA GTT GGG GAA CTA ACT AAT CCG
          lys cys val lys arg glu his ala gln gly leu val leu val gly glu leu thr asn pro

2161/699  CCT TAC CAC GAG CTG GCA TAC GAA GGA TTA CGT ACA CGA CCC GCT GCC CCT TAC CAT ATC
          pro tyr his glu leu ala tyr glu gly leu arg thr arg pro ala ala pro tyr his ile
```

Figure 4b. See legend on last page of this sequence

```
2221/719  GAA ACA CTG GGG GTC ATT GGA ACA CCG GGG TCA GGT AAG TCG GCC ATC ATA AAA TCT ACG
          glu thr leu gly val ile gly thr pro gly ser gly lys ser ala ile ile lys ser thr

2281/739  GTA ACA CTA AAA GAC CTC GTA ACT AGC GGT AAG AAA GAA AAT TGC AAA GAA ATA GAG AAT
          val thr leu lys asp leu val thr ser gly lys lys glu asn cys lys glu ile glu asn

2341/759  GAC GTC CAG AAA ATG CGG GGA ATG ACT ATA GCT ACG AGA ACG GTA GAC TCG GTA CTT CTT
          asp val gln lys met arg gly met thr ile ala thr arg thr val asp ser val leu leu

2401/779  AAT GGA TGG AAG AAA GCA GTA GAC GTC CTA TAT GTG GAT GAA GCG TTT GCA TGT CAT GCA
          asn gly trp lys lys ala val asp val leu tyr val asp glu ala phe ala cys his ala

2461/799  GGC ACC TTA ATG GCA TTG ATT GCC ATT GTC AAA CCG AGA CGT AAA GTA GTA CTG TGC GGC
          gly thr leu met ala leu ile ala ile val lys pro arg arg lys val val leu cys gly

2521/819  GAC CCG AAG CAG TGG CCC TTC TTT AAT TTA ATG CAA CTG AAG GTA AAC TTC AAC AAC CCC
          asp pro lys gln trp pro phe phe asn leu met gln leu lys val asn phe asn asn pro

2581/839  GAG CGA GAC CTG TGT ACT TCC ACC CAT TAT AAA TAT ATC TCT CGC AGG TGC ACC CAA CCT
          glu arg asp leu cys thr ser thr his tyr lys tyr ile ser arg arg cys thr gln pro

2641/859  GTT ACA GCC ATA GTG TCT ACA TTA CAC TAT GAC GGA AAG ATG AGG ACT ACG AAT CCC TGC
          val thr ala ile val ser thr leu his tyr asp gly lys met arg thr thr asn pro cys

2701/879  AAA AGG GCT ATC GAA ATA GAC GTA AAC GGA TCG ACT AAG CCC AAG AAA GGA GAC ATA GTG
          lys arg ala ile glu ile asp val asn gly ser thr lys pro lys lys gly asp ile val

2761/899  TTG ACG TGT TTC CGT GGG TGG GTT AAG CAG GGG CAA ATC GAT TAC CCC GGA CCC GGA GGT
          leu thr cys phe arg gly trp val lys gln gly gln ile asp tyr pro gly pro gly gly

2821/919  CAT GAC CGT GCA GCT TCT CAA GGG CTA ACC AGA AGG GGC GTT TAT GCG GTC AGA CAG AAA
          his asp arg ala ala ser gln gly leu thr arg arg gly val tyr ala val arg gln lys

2881/939  GTA AAT GAA AAC CCA CTA TAT GCA GAG AAG TCA GAA CAC GTT AAC GTG TTA CTT ACT AGG
          val asn glu asn pro leu tyr ala glu lys ser glu his val asn val leu leu thr arg

2941/959  ACG GAA GAT CGC ATA GTG TGG AAG ACA CTG CAA GGG GAT CCT TGG ATT AAG TAC CTC ACT
          thr glu asp arg ile val trp lys thr leu gln gly asp pro trp ile lys tyr leu thr

3001/979  AAC GTT CCA AAA GGG AAC TTT ACA GCC ACT TTA GAA GAA TGG CAG GCG GAA CAC GAG GAC
          asn val pro lys gly asn phe thr ala thr leu glu glu trp gln ala glu his glu asp

3061/999  ATT ATG AAG GCC ATT AAT TCT ACA TCC ACA GTA TCT GAC CCT TTC GCC AGC AAA GTG AAT
          ile met lys ala ile asn ser thr ser thr val ser asp pro phe ala ser lys val asn

3121/1019 ACA TGC TGG GCT AAA GCT ATT ATA CCC ATC CTA AGA ACG GCA GGG ATA GAA CTT ACA TTC
          thr cys trp ala lys ala ile ile pro ile leu arg thr ala gly ile glu leu thr phe

3181/1039 GAG CAG TGG GAA GAT CTA TTC CCG CAA TTT CGT AAT GAC CAA CCT TAC TCC GTG ATG TAT
          glu gln trp glu asp leu phe pro gln phe arg asn asp gln pro tyr ser val met tyr

3241/1059 GCC CTA GAT GTG ATA TGT ACC AAG ATG TTC GGC ATG GAT CTG AGC AGT GGG ATC TTC TCT
          ala leu asp val ile cys thr lys met phe gly met asp leu ser ser gly ile phe ser

3301/1079 CGT CCT GAG ATA CCT CTA ACG TTC CAT CCC GCG GAC GTC GGC CGA GTG AGA GCT CAC TGG
          arg pro glu ile pro leu thr phe his pro ala asp val gly arg val arg ala his trp
```

Figure 4c. See legend on last page of this sequence

```
3361/1099 GAT AAC TCC CCA GGA GGG CAG AAG TTT GGG TAT AAC AAG GCG GTA ATC CCA ACT TGC AAG
          asp asn ser pro gly gly gln lys phe gly tyr asn lys ala val ile pro thr cys lys

3421/1119 AAA TAC CCA GTG TAC TTA AGA GCA GGA AAA GGG GAC CAA ATA CTC CCC ATA TAT GGC AGA
          lys tyr pro val tyr leu arg ala gly lys gly asp gln ile leu pro ile tyr gly arg

3481/1139 GTT TCA GTC CCA TCG GCA CGG AAC AAT TTA GTT CCC TTA AAC AGA AAT CTA CCA CAC TCG
          val ser val pro ser ala arg asn asn leu val pro leu asn arg asn leu pro his ser

3541/1159 CTA ACT GCA AGC CTG CAG AAA AAA GAA GCA GCT CCC TTG CAC AAG TTC CTT AAC CAA CTA
          leu thr ala ser leu gln lys lys glu ala ala pro leu his lys phe leu asn gln leu

3601/1179 CCA GGA CAC AGT ATG CTG CTG GTC TCT AAG GAA ACA TGC TAT TGC GTG TCC AAG CGA ATC
          pro gly his ser met leu leu val ser lys glu thr cys tyr cys val ser lys arg ile

3661/1199 ACA TGG GTC GCT CCG CTG GGA GTC AGA GGA GCT GAC CAC AAC CAT GAC CTG CAT TTC GGG
          thr trp val ala pro leu gly val arg gly ala asp his asn his asp leu his phe gly

3721/1219 TTC CCA CCA CTG TCC AGA TAC GAC CTT GTG GTG GTT AAT ATG GGA CAA CCG TAC AGG TTC
          phe pro pro leu ser arg tyr asp leu val val val asn met gly gln pro tyr arg phe

3781/1239 CAT CAC TAC CAG CAG TGC GAG GAG CAT GCC GGC CTC ATG AGG ACG TTG GCC CGG TCA GCA
          his his tyr gln gln cys glu glu his ala gly leu met arg thr leu ala arg ser ala

3841/1259 CTC AAC TGC CTA AAA CCA GGA GGA ACA TTA GCC CTG AAA GCA TAT GGT TTC GCC GAC TCC
          leu asn cys leu lys pro gly gly thr leu ala leu lys ala tyr gly phe ala asp ser

3901/1279 AAT AGT GAG GAC GTT GTT CTG TCT TTA GCG AGG AAA TTC GTG CGG GCA TCC GCA GTG AGA
          asn ser glu asp val val leu ser leu ala arg lys phe val arg ala ser ala val arg

3961/1299 CCA TCG TGT ACA CAG TTT AAC ACA GAG ATG TTC TTT GTA TTT AGG CAG CTG GAC AAC GAT
          pro ser cys thr gln phe asn thr glu met phe phe val phe arg gln leu asp asn asp

4021/1319 CGT GAG CGC CAA TTC ACT CAG CAT CAC TTG AAT TTA GCA GTA TCC AAT ATA TTC GAC AAT
          arg glu arg gln phe thr gln his his leu asn leu ala val ser asn ile phe asp asn

4081/1339 TAT AAA GAC GGA TCC GGA GCA GCT CCT TCT TAT CGC GTT AAG AGA ATG AAT ATC GCA GAC
          tyr lys asp gly ser gly ala ala pro ser tyr arg val lys arg met asn ile ala asp

4141/1359 TGC ACA GAA GAA GCA GTG GTG AAC GCA GCT AAC GCG CGG GGA AAA CCT GGG GAC GGA GTA
          cys thr glu glu ala val val asn ala ala asn ala arg gly lys pro gly asp gly val

4201/1379 TGC AGA GCT ATC TTC AAA AAG TGG CCG AAG TCA TTT GAG AAC GCT ACC ACT GAA GTG GAA
          cys arg ala ile phe lys lys trp pro lys ser phe glu asn ala thr thr glu val glu

4261/1399 ACC GCG GTC ATG AAA CCA TGC CAC AAC AAG GTT GTT ATA CAT GCA GTG GGT CCT GAT TTT
          thr ala val met lys pro cys his asn lys val val ile his ala val gly pro asp phe

4321/1419 AGA AAG TAC ACG TTG GAG GAA GCG ACG AAG CTA CTG CAG AAC GCA TAC CAT GAT GTG GCA
          arg lys tyr thr leu glu glu ala thr lys leu leu gln asn ala tyr his asp val ala

4381/1439 AAG ATA GTG AAC GAG AAA GGC ATC TCC TCG GTA GCT ATA CCG CTG CTC TCA ACA GGT ATC
          lys ile val asn glu lys gly ile ser ser val ala ile pro leu leu ser thr gly ile

4441/1459 TAT GCT GCC GGA GCT GAT CGC CTG GAT CTC TCG CTG AGA TGT CTT TTC ACC GCG CTG GAT
          tyr ala ala gly ala asp arg leu asp leu ser leu arg cys leu phe thr ala leu asp
```

Figure 4d. See legend on last page of this sequence

```
4501/1479  CGT ACG GAT GCG GAT GTC ACA ATA TAT TGC CTA GAT AAG AAG TGG GAG CAA CGC ATA GCA
           arg thr asp ala asp val thr ile tyr cys leu asp lys lys trp glu gln arg ile ala

4561/1499  GAT GCT ATT AGG ATG CGA GAA CAA GTA ACT GAA TTA AAA GAT CCG GAC ATA GAG ATA GAT
           asp ala ile arg met arg glu gln val thr glu leu lys asp pro asp ile glu ile asp

4621/1519  GAA GGA TTA ACC CGG GTA CAC CCA GAT AGC TGC CTC AAG GAT CAC ATA GGC TAC AGT ACC
           glu gly leu thr arg val his pro asp ser cys leu lys asp his ile gly tyr ser thr

4681/1539  CAG TAT GGG AAA TTG TAC TCA TAC TTT GAA GGT ACT AAA TTC CAC CAA ACC GCA AAA GAC
           gln tyr gly lys leu tyr ser tyr phe glu gly thr lys phe his gln thr ala lys asp

4741/1559  ATA GCC GAG ATT CGT GCG CTG TTT CCT GAT GTA CAA GCC GCT AAC GAA CAA ATC TGC CTG
           ile ala glu ile arg ala leu phe pro asp val gln ala ala asn glu gln ile cys leu

4801/1579  TAC ACT TTA GGC GAA CCG ATG GAG TCC ATA CGC GAA AAG TGC CCA GTC GAA GAC TCC CCG
           tyr thr leu gly glu pro met glu ser ile arg glu lys cys pro val glu asp ser pro

4861/1599  GCA TCA GCA CCT CCT AAG ACA ATA CCT TGC CTA TGT ATG TAT GCT ATG ACA GCC GAA CGT
           ala ser ala pro pro lys thr ile pro cys leu cys met tyr ala met thr ala glu arg

4921/1619  ATT TGC CGC GTA CGC AGT AAC TCC GTA ACG AAC ATA ACG GTG TGC TCA TCC TTT CCG TTA
           ile cys arg val arg ser asn ser val thr asn ile thr val cys ser ser phe pro leu

4981/1639  CCC AAG TAC CGA ATA AAG AAC GTA CAA AAG ATA CAG TGC ACG AAA GTG
           pro lys tyr arg ile lys asn val gln lys ile gln cys thr lys val
```

Figure 4 Translated sequence of Aura virus. This sequence starts near the 5'terminus of the genome, although the exact 5' end is not known. The translated sequence shown encompasses nsP1, nsP2, and the N-terminal (conserved) region of nsP3. Nucleotides are numbered from the beginning of the sequence; amino acids are numbered from the beginning of the open reading frame.

```
 23 MEKPTVHVDVDPQSPFVLQLQKSFPQFEIVAQQVTPNDHANARAFSHLAS  72
    ||||.|:||||||||||:||||||||||:||||||||||||||||||||
  1 MEKPVVNVDVDPQSPFVVQLQKSFPQFEVVAQQVTPNDHANARAFSHLAS  50

 73 KLIEHEIPTSVTILDIGSAPARRMYSEHKYHCVCPMRSPEDPDRLMNYAS 122
    |||| |:||..|||||||||||||:|||.|||||||||||||||:|.|||
 51 KLIELEVPTTATILDIGSAPARRMFSEHQYHCVCPMRSPEDPDRMMKYAS 100

123 RLADKAGEITNKRLHDKLADLKSVLESPDAETGTICFHNDVICRTTAEVS 172
    :||:||..||||.||:|: ||:.||:.||||||..:||||||.|.  || |
101 KLAEKACKITNKNLHEKIKDLRTVLDTPDAETPSLCFHNDVTCNMRAEYS 150

173 VMQNVYINAPSTIYHQALKGVRKLYWIGFDTTQFMFSSMAGSYPSYNTNW 222
    |||:||||||:||||||:||||.|||||||||||||.||||||.||||||
151 VMQDVYINAPGTIYHQAMKGVRTLYWIGFDTTQFMFSAMAGSYPAYNTNW 200

223 ADERVLEARNIGLCSTKLREGTMGKLSTFRKKALKPGTNVYFSVGSTLYP 272
    |||:|||||||||||||.||  ||||.:|||.||||..||||||||||||
201 ADEKVLEARNIGLCSTKLSEGRTGKLSIMRKKELKPGSRVYFSVGSTLYP 250

273 ENRADLQSWHLPSVFHLKGKQSFTCRCDTAVNCEGYVVKKITISPGITGR 322
    |:||.|||||||||||.||||:||||||.|.|||||||||||||||||||
251 EHRASLQSWHLPSVFHLNGKQSYTCRCDTVVSCEGYVVKKITISPGITGE 300

323 VNRYTVTNNSEGFLLCKITDTVKGERVSFPVCTYIPPSICDQMTGILATD 372
    .  |.||:|||||||||:|||||||||||||||||:.|||||||:|||
301 TVGYAVTHNSEGFLLCKVTDTVKGERVSFPVCTYIPATICDQMTGIMATD 350

373 IQPEDAQKLLVGLNQRIVVNGKTNRNTNTMQNYLLPAVATGLSKWAKERK 422
    | |:||||||||||||||:||:||||||||||||| :| |:||||||||
351 ISPDDAQKLLVGLNQRIVINGRTNRNTNTMQNYLLPIIAQGFSKWAKERK 400

423 ADCSDEKPLNVRERKLAFGCLWAFKTKKIHSFYRPPGTQTIVKVAAEFSA 472
    .| .:|| |..||||||.:||||||:|||:||||||||||.|||:|.|||
401 DDLDNEKMLGTRERKLTYGCLWAFRTKKVHSFYRPPGTQTCVKVPASFSA 450

473 FPMSSVWTTSLPMSLRQKVKLLLVKKTNKPVVTITDTAVKNAQEAYNEAV 522
    |||||||||||||||||||:|| | .|.:.:.:: :.:. |.:|..|:::|
451 FPMSSVWTTSLPMSLRQKLKLALQPKKEEKLLQVSEELVMEAKAAFEDAQ 500

523 ETAEAEEKAKALPPLKP.TAPPVAEDVKCEVTDLVDDAGAALVETPRGKI 571
    |.| ||.  .|||||| : .: ..|.:| |||.:| .| ||||||||||.:
501 EEARAEKLREALPPLVADKGIEAAAEVVCEVEGLQADIGAALVETPRGHV 550
```

Figure 5a See legend on last page of this sequence.

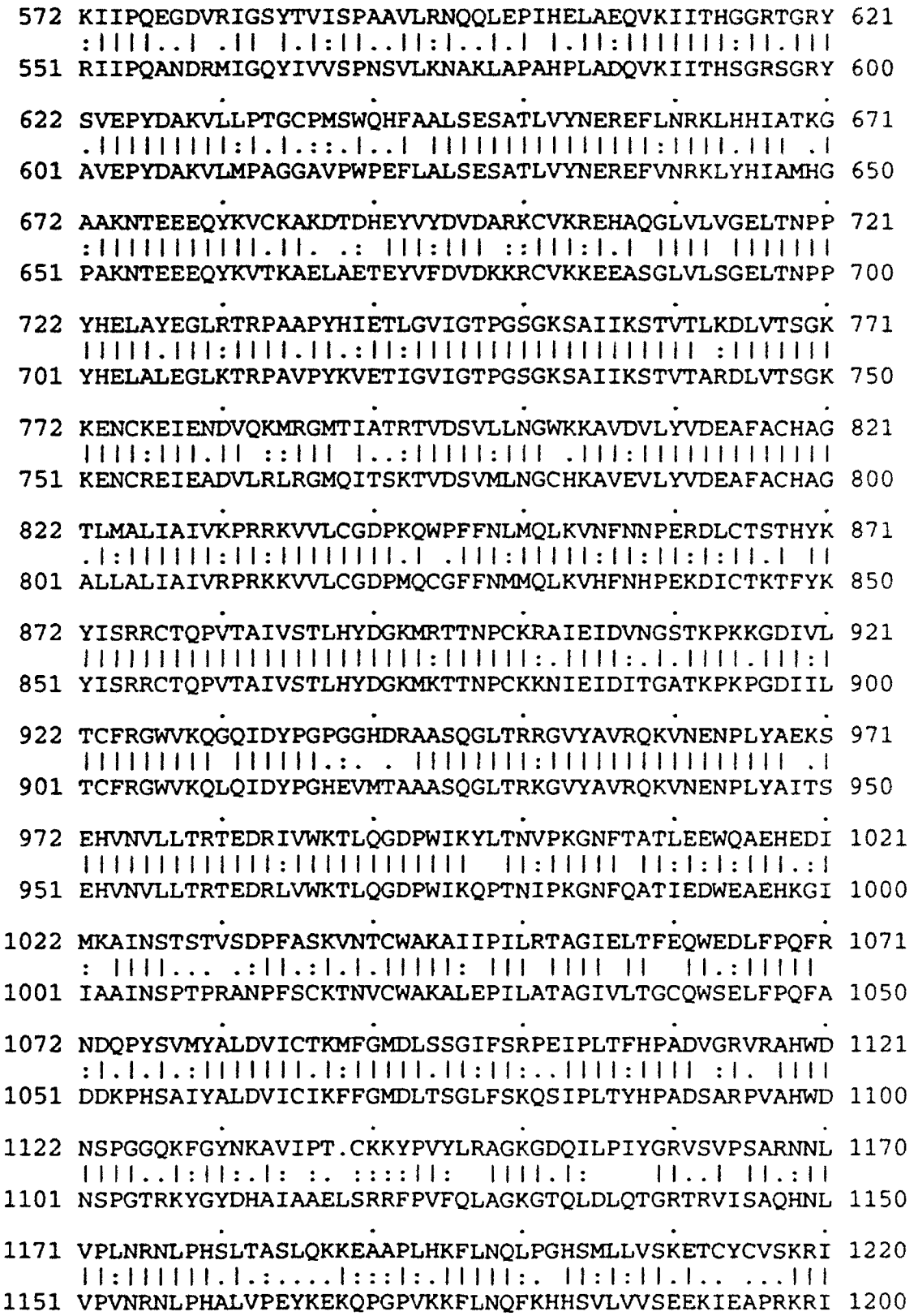

Figure 5b continued, see legend on next page.

Figure 5, continued. Alignment of the deduced amino acid sequences of Aura virus (top line) and Sindbis virus (lower line) in the region encoding nsP1, nsP2, and the N-terminal (conserved) domain of nsP3. Amino acid identities are indicated with solid vertical lines; dots indicate functionally similar residues.

Sindbis-like virus in the Americas We have previously shown that Western equine encephalitis virus, previously thought to be closely related to Sindbis virus, is in fact a recombinant virus in which most of the genome was derived from Eastern equine encephalitis virus and only the surface glycoproteins were derived from a Sindbis-like virus (Hahn et al., 1988). Furthermore, Western equine encephalitis virus lacks the characteristic Sindbis 3' nontranslated region.

Aura virus is widely distributed in South America, having been isolated in Brazil and in Northern Argentina. Analysis of the data is not yet complete, but it is possible that Aura virus represents the ancestral Sindbis-like virus, and that it was transmitted to the Old World to serve as the founder of the Sindbis viruses in the Old World, as we previously postulated (Levinson et al., 1990). Aura virus may have served as one of the parents of Western equine encephalitis virus, contributing its glycoproteins to this recombinant virus (Hahn et al., 1988).

Conclusions

The Sindbis-like viruses, which are found throughout the Old World from Northern Europe to Africa, India, the Philippines and the Australasian region including New Guinea, are a clearly identifiable group of viruses. They share a minimum of 80% amino acid sequence identity in nsP2 and possess a characteristic and conserved 3' nontranslated region. It is of considerable interest that viruses belonging to this group coexist in many parts of the world with other alphaviruses that are demonstrably different in their epidemiology, serology, organization of the 3' nontranslated region, and evolutionary history, even though most of these non-Sindbis alphaviruses cause diseases very similar to those caused by the Sindbis-like viruses.

We have clearly shown that high throughput automated DNA sequencing is ideally suited to the rapid analysis of an RNA virus family such as the

alphaviruses. These procedures are rapid and generate large amounts of useful information very quickly. Such procedures would be very useful in defining the origin and spread of an epidemic virus.

We have shown that Aura virus is a New World representative of the Sindbis viruses. Further analysis is required to determine whether it is one of the parents of Western equine encephalitis virus, but the hypothesis that Western equine encephalitis virus is an emergent virus that arose by recombination has received further support from these studies.

References

Griffin, D. E. (1986). Alphavirus pathogenesis and immunity. In "The Togaviridae and Flaviviridae" (S. Schlesinger and M. J. Schlesinger, Ed.), pp. 209-250. Plenum Publishing Corp., New York.

Gubler, U., and Hoffman, B. J. (1983). A simple and very efficient method for generating cDNA libraries. Gene 25, 263-269.

Hahn, C. S., Lustig, S., Strauss, E. G., and Strauss, J. H. (1988). Western equine encephalitis virus is a recombinant virus. Proc. Natl. Acad Sci. USA 85, 5997-6001.

Hahn, C. S., Strauss, E. G., and Strauss, J. H. (1985). Sequence analysis of three Sindbis virus mutants temperature-sensitive in the capsid autoprotease. Proc. Natl. Acad. Sci. USA 82, 4648-4652.

Levinson, R., Strauss, J. H., and Strauss, E. G. (1990). Determination of the complete nucleotide sequence of the genomic RNA of O'Nyong-nyong virus and its use in the construction of phylogenetic trees. Virology 175, 110-123.

Rice, C., M., Lenches, E. M., Eddy, S. R., Shin, S. J., Sheets, R. L., and Strauss, J. H. (1985). Nucleotide sequence of yellow fever virus: Implications for flavivirus gene expression and evolution. Science 229, 726-733.

Rice, C. M., and Strauss, J. H. (1981). Synthesis, cleavage, and sequence analysis of DNA complementary to the 26S messenger RNA of Sindbis virus. J. Mol. Biol. 150, 315-340.

Sambrook, J., Fritsch, E. F., and Maniatis, T. (1989). Molecular Cloning: A Laboratory Manual, Second, Cold Spring Harbor Laboratory Press, Cold Spring Harbor.

Shirako, Y., Niklasson, B., Dalrymple, J. M., Strauss, E. G., and Strauss, J. H. (1991). Structure of the Ockelbo virus genome and its relationship to other Sindbis viruses. Virology 182, 753-764.

Strauss, E. G., Rice, C. M., and Strauss, J. H. (1984). Complete nucleotide sequence of the genomic RNA of Sindbis virus. Virology 133, 92-110.

Strauss, E. G., Stec, D. S., Schmaljohn, A. L., and Strauss, J. H. (1991). Identification of antigenically important domains in the glycoproteins of Sindbis virus by analysis of antibody escape variants. J. Virol. in press,

Wang, K.-S., and Strauss, J. H. (1991). Use of a lgt11 expression library to localize a neutralizing antibody-binding site in glycoprotein E2 of Sindbis virus. J. Virol. 65, 7037-7040.